Hello, who is this? The relationship between linguistic and speaker-dependent information in the acoustics of consonants

Smorenburg, B.J.L.

CHAPTER 4

Effects of the landline telephone filter

Abstract

Previous work on telephone speech investigating effects of phonetic context and syllabic position on acoustics and speaker variation found different effects for Dutch fricatives /x/ and /s/ (Smorenburg & Heeren, 2020). This was attributed to the narrowband telephone filter cutting of spectral energy from /s/, not /x/. Using English data that was simultaneously recorded as broadband and telephone speech, this work shows that linguistic effects are affected by the telephone filter. Additionally, linguistic context effects on speaker variation again show

that fricatives in labial contexts contain more between-speaker variation than fricatives in non-labial contexts. However, this was only the case for following labial context, not preceding labial context, and no substantial difference was found between /s/ in coda and onset position.

This chapter has been submitted and parts of this chapter have been presented at:

Smorenburg, L., & Heeren, W. (2021). Effects of speech channel on acoustic measurements and speaker discrimination from /s/. In *29th conference of IAFPA*. Marburg, Germany: University of Marburg.

Smorenburg, L., & Heeren, W. (2022). The effects of linguistic contexts on the acoustics and strength-of-evidence of /s/. In *30th conference of IAFPA* (pp. 13–14). Prague, Czech Republic: Charles University.

4.1 Introduction

Social and idiosyncratic information in speech play a large role in everyday communication. Perception studies have for instance shown that sentence interpretation is dependent on (inferred) speaker information (Van Berkum, Van Den Brink, Tesink, Kos, & Hagoort, 2008). Speech acoustics can be used to characterize individual speakers and in forensic speaker comparisons (FSC), the idiosyncratic information in voices is analyzed, and may serve as evidence in court. To improve FSC, researchers have been trying to establish what factors, both linguistic and extra-linguistic, affect the idiosyncratic information in speech.

Different speech segments hold different amounts of idiosyncratic information. Namely, vowels typically contain more speaker information than consonants (e.g., Van den Heuvel, 1996), although see Schindler and Draxler (2013). Amongst the consonants, nasals and fricatives contain more speaker information than other consonants (Kavanagh, 2012; Van den Heuvel, 1996). Moreover, there is some evidence that the same segment might also contain slightly different amounts of speaker information in different linguistic contexts or positions (e.g., see Heeren, 2020a on word class; McDougall, 2004 on lexical stress; Smorenburg & Heeren, 2020 and Su, Li, & Fu, 1974 on phonetic context and idiosyncrasies in coarticulation). On the one hand, some linguistic contexts and positions may result in lower within-speaker variation which may serve to increase speaker-specificity, for example in content words (Heeren, 2020a) and stressed vowels (McDougall, 2004). On the other hand, the degree and timing of coarticulatory movements and reduction may be specific to speakers (cf. Nolan, 1983, Chapter 3), thus increasing between-speaker variation (Smorenburg & Heeren, 2020; Su, Li, & Fu, 1974).

One major concern in FSC is the effects of telephone filters on speaker discrimination. In the Netherlands, wiretapped telephone conversations are common in FSC and it is therefore relevant to know

how telephone filters affect speech acoustics and speaker discrimination. Although the effects of telephone filters on speech acoustics have previously been investigated for some vowels (Byrne & Foulkes, 2004; Künzel, 2001), less is known about their effect on consonants. Given that some consonants, such as sibilant fricatives, have their spectral peak at frequencies outside of the upper limit of most telephone filters, the effect of telephone filters may be high for some consonants. In fact, it has been observed that fricative discrimination in narrowband telephone signals can be difficult (Bessette et al., 2002). Sibilant fricative /s/ in particular has a spectral center of gravity above 7 kHz in some groups of speakers (Munson, McDonald, DeBoe, & White, 2006). Given that /s/ acoustics can convey some information about speaker identity, the telephone filter is expected to have an effect on the idiosyncratic information in /s/.

Previous research on fricatives /s/ and /x/ showed that /s/ still contained significant amounts of idiosyncratic information, even in a landline telephone bandpass of 300 – 3,400 Hz (Smorenburg & Heeren, 2020). Dutch /s/, however, has lower-frequency spectral characteristics than English /s/, which could mean that less idiosyncratic information is available for English /s/ in narrowband signals. Spectral characteristics from fricatives are furthermore strongly affected by labial coarticulation (e.g., Koenig, Shadle, Preston, & Mooshammer, 2013; Munson, 2004), which seemed to affect the speaker-specificity of Dutch fricatives in systematic ways (Smorenburg & Heeren, 2020). The current work investigated effects of linguistic context on the acoustics and speaker variation of British English /s/, also considering effects of and interactions with the landline telephone filter. Although the signal characteristics of landline signals are not entirely representative of the mobile signals that are commonly used in modern communications, the band pass of landline filters is still relevant in the forensic context.

4.1.1 Fricative /s/ acoustics

The alveolar fricative /s/ is articulated by making a narrow constriction at the alveolar ridge. This creates a turbulent airflow which results in an acoustic signal with aperiodic frication noise (Stevens, 2000). This frication noise predominantly reflects the resonance characteristics of the anterior cavity, which, for /s/, is the space between the alveolar constriction and the lips (Stevens, 2000). The smaller that space, the higher the frequency of the frication noise. The alveolar sibilant /s/, for example, has higher-frequency frication noise than post-alveolar /ʃ/ (e.g., Jongman, Wayland, & Wong, 2000). This difference in anterior cavity size is also reflected in effects of sex; male speakers generally have a larger vocal tract and thus lower /s/ frequencies than female speakers (Li et al., 2016; Schwartz, 1968). Cross-linguistic differences have also been attested. Speakers of Dutch, e.g., have laminal articulations of /s/ where the constriction is made with the tongue front/blade. This is different for speakers of English or French where the constriction is apical, i.e., made with a pointed tongue tip. As a result, the anterior cavity in /s/ articulation is larger for speakers of Dutch, resulting in a lower center of gravity in Dutch than in English (Collins & Mees, 1984; Quené, Orr, & Van Leeuwen, 2017). Considering the differences in phoneme inventories and articulatory settings, there are some potentially relevant differences between English and Dutch. For example, it has been observed that Dutch generally has more muscular tension in the lips, whereas in British English the lips are less active, resulting in the stereotype of a 'stiff upper lip' (cf. Collins & Mees, 1984). This goes hand in hand with the vowel inventory: Dutch has more rounded vowels than English, which can be front or back, whereas English round vowels are all back. This is relevant for the effect of phonetic context in this work, as lengthening of the anterior cavity can be achieved by both protruding the lips or having a more posterior tongue constriction in fricative articulation.

Phonetic context also affects the size of the anterior cavity; protruding the lips in anticipatory lip-rounding lengthens the anterior cavity and lowers the frication noise (e.g., Koenig, Shadle, Preston, & Mooshammer, 2013; Munson, 2004; Shadle & Scully, 1995). Another linguistic effect that influences fricative acoustics is syllabic position, although there are contradicting reports, specifically for /s/. Generally

speaking, consonants in coda position are articulated with less effort than consonants in onset position (Ohala & Kawasaki, 1984). For fricatives, coda reduction is observed for fricatives in general but not consistently across temporal and spectral measurements for /s/ in particular (Cunha & Reubold, 2015; Redford & Diehl, 1999; Solé, 2003).

Previous research has shown that there can be cross-linguistic differences in patterns of coarticulation. Most generally, it has been hypothesized that languages can be characterized by the direction of coarticulation. For example, it has been claimed that French shows predominantly anticipatory coarticulation, whereas English shows predominantly carry-over coarticulation (Hoole, Nguyen-Trong, & Hardcastle, 1993). However, acoustic evidence only varyingly corroborates this hypothesis. For example, Magen (1997) found no evidence for more carry-over than anticipatory V-V coarticulation in English. Niebuhr, Clayards, Meunier, and Lancia (2011), on the other hand, found that sibilant sequences in English show exclusively carry-over place articulation, whereas French showed both carry-over and anticipatory place assimilation. Looking at labial coarticulation specifically, many studies show generally large effects of anticipatory labialization in English (e.g., Bell-Berti & Harris, 1982; Koenig et al., 2013; Munson, 2004; Nittrouer & Whalen, 1989; Soli, 1981). Not many studies focus on carry-over labialization in English, although some studies do investigate the combined effect of carry-over and anticipatory labialization in VCV sequences (e.g., Shadle & Scully, 1995). Due to these crosslinguistic differences, it is possible that previous findings on the context-dependency of speaker variation in Dutch /s/ do not generalize to English.

4.1.2 Idiosyncratic information in /s/

Amongst the consonants, nasals and fricatives seem to contain the highest amounts of idiosyncratic information. Nasals are often reported to be robust to many contextual influences and therefore show relatively little within-speaker variation, which makes them relatively speaker specific

(Rose, 2002). Fricatives, particularly /s/, also carry social information about the speaker and therefore have relatively high between-speaker variation, which also makes them relatively speaker-specific. Regarding the between-speaker variation in fricatives, it has been shown that social class and gender significantly affect /s/ productions (Stuart-Smith, 2007) and that even sexual orientation is encoded in and perceived from the acoustics of /s/ (Munson et al., 2006; Tracy, Bainter, & Satariano, 2015). For speakers of Dutch, /s/ acoustics have also been shown to contain information about ethnicity (Ditewig, Smorenburg, Quené, & Heeren, 2021) and region (Ditewig, Pinget, & Heeren, 2019). These social and linguistic variables, along with the acoustic reflection of the speaker's vocal tract size, all contribute to this sound being relatively speaker-specific, which makes it a potentially useful sound in FSC.

There also seem to be differences in the amount of idiosyncratic information within speech sounds that can be related to prosodic structure and phonetic context. Regarding prosodic structure, it seems that speech articulated with more effort is more precise and therefore more consistent within (and also between) speakers. For example, content words seem to contain slightly more speaker information than function words (Heeren, 2020a) and stressed vowels seem to contain slightly more speaker information than unstressed vowels (McDougall, 2004). Conversely, less articulatory effort allows for more freedom in reduced forms. He, Dellwo and colleagues studied between-speaker variation in intensity and formant contours of syllables and found more variation in the second half of syllables, i.e., towards the syllable coda (He & Dellwo, 2017; He, Zhang, & Dellwo, 2019). This was explained by the relative articulatory freedom of codas, whereas realizations of onsets are more constrained. It has also been observed that idiosyncrasies exist in coarticulation (cf. Nolan, 1983, Chapter 3). With regards to /s/, fricative realizations are highly dependent on contextual labialization and /s/ in labial contexts generally showed slightly more between-speaker variation than fricatives in non-labial contexts (Smorenburg & Heeren, 2020). Similarly, this was shown for nasal consonants /n/ and /m/ in contexts with coarticulation (e.g., Su, Li, & Fu, 1974).

4.1.3 Telephone signals and telephone speech

Speech transmitted over telephones loses acoustic information due to the limited band passes used in telephony. Telephone signals can be subdivided into two main types; landline and mobile signals. Landline telephone signals have a narrow band pass of about 300 – 3,400 Hz, meaning that spectral energy below 300 and above 3,400 Hz is strongly attenuated or lacking altogether (Künzel, 2001). Although some mobile signals have a very similar band pass to landline signals, the signal is much less stable. For example, the Adaptive Multi-Rate (AMR) narrowband codec (the compression technology used in 2G and 3G signals) that was standardized for the Global System Mobile Communication (GSM) network has a similar band pass of 200 – 3,400 Hz (Bessette et al., 2002). However, its bit rates can change rapidly, which can lower the upper frequency cut-off from 3,400 Hz to 2,800 Hz (Guillemin & Watson, 2006). More modern cellular technology uses much wider bandwidths, e.g., the Adaptive Multi-Rate Wideband (AMR-WB) codec used in 4G signals covers a 50 – 7,000 Hz band pass and thus provides better fricative differentiation (Bessette et al., 2002). For speech sounds with high-frequency characteristics such as /s/, this upper cut-off captures more information than landline signals and mobile predecessors. However, the AMR-WB still has a varying bit rate depending on channel conditions; the signal changes to half-rate when channel conditions are considered good based on harmonics-to-noise ratios (Bessette et al., 2002).

In the Netherlands, telephony providers are legally required to make wiretapping available for both landline and mobile telephone signals (Van de Pol, 2006). When a call is wiretapped, an authorized third party can listen in on the call and record it. Such recordings may be processed for police investigations. As a result, much of the speech material in forensic casework consists of wiretapped telephone conversations.

Effects of telephone signals on speech can be both signal-related and behavioral in nature. Signal-related effects have mostly been described for vowels in landline signals; vowel formants that are situated

near the lower telephone cut-off are affected in landline signals (Künzel, 2001) and in mobile signals (Byrne & Foulkes, 2004). Specifically, the measurements of F1 values might shift upward. In automatic speaker recognition, which uses more holistic speech features such as Mel-frequency cepstral coefficients, mismatches in speech channel also have significant effects on speaker discrimination when it concerns telephone versus studio recordings (Van der Vloed, Kelly, & Alexander, 2020). For auditory-acoustic analysis however, where linguistic-phonetic speech features are examined and which is more common in forensic casework across the globe (Gold & French, 2011, 2019), it is not yet clear what effects different kind of telephone filters may have on consonants in particular. Some previous research has attempted to replicate telephone filters by using a 500 – 4,000 Hz frequency range for extracting measurements from /s/ in broadband signals (Kavanagh, 2012). Using discriminant analysis, Kavanagh (2012) found similar speaker-classification accuracies for the simulated telephone filtering condition compared to a broadband condition (500-8,000 Hz). When using likelihood-ratio testing, however, better speaker classifications were obtained in the narrowband compared to the broadband filtering condition, which Kavanagh remarked was notable. As will become clear in the current work, telephone signals are not only different from broadband signals in their frequency range, but generally show different spectral shapes due to noise and compression mechanisms in the telephone codec. It is therefore necessary to use actual telephone signals to test the effect of telephone filters on /s/.

Regarding behavioral effects, a speaker's "telephone voice" is often subject to the Lombard effect, i.e., the increase of vocal effort in the presence of noise (Junqua, Fincke, & Field, 1999). When conversing over the telephone speakers cannot be seen by the listener, meaning that hand gestures and facial expressions cannot be used and acoustic means might replace them. In a study on the use of intonation in turn-taking in telephone versus face-to-face conversations, differences were found in speakers' pitch ranges, where a larger pitch range was associated with holding the turn in face-to-face conversation but with changing the turn in telephone conversation (Oliveira & Freitas, 2008). Although perception results subsequently showed that intonation alone did not

seem a sufficient cue for turn transition, this study confirms that speakers display different uses of intonation in their production across speech conditions.

To summarize, telephone speech behavior may differ from other speech behavior and telephone signals are limited in their frequency range, which can affect fricative discrimination. It is not yet clear how the loss of acoustic information may impact the speaker information in /s/, although some research has shown that a limited frequency range does not necessarily lead to decreased speaker classification for /s/.

4.1.4 Research questions

This study investigated the effects of the telephone filter and of phonetic context and syllabic position on the acoustics and speaker characteristics of /s/. Previous research has shown that acoustic-phonetic features from Dutch /s/ still contain significant amounts of idiosyncratic information in landline telephone recordings (Smorenburg & Heeren, 2020). That speech corpus, however, only contained telephone signals and, therefore, did not allow for a direct comparison between telephone and studio channels (i.e., high-quality recordings). Here, we investigate the effect of the landline telephone filter on /s/ in direct comparison with simultaneously recorded studio speech in British English data from the West Yorkshire Regional English Database (WYRED; Gold, Ross, & Earnshaw, 2018). The acoustics of /s/ may be assumed to be highly affected by the telephone wiretapping because its spectral peak falls outside of the telephone band. However, it is possible that the between-speaker variation in spectral peak values is also (partly) reflected in the weaker spectral energy at lower frequencies. Moreover, some acoustic-phonetic measurements might be more robust to telephone filters than others.

Additionally, previous research showed that phonetic context and syllabic position affect the acoustics and speaker information in fricatives from Dutch landline telephone speech (Smorenburg & Heeren, 2020). This work further investigated the possible interactions between

linguistic effects and signal bandwidth and the generalizability of previous Dutch results across languages. It is predicted that English /s/ will show effects of both contextual labialization and syllabic position. Based on the hypothesis that English is a carry-over language, it is expected that carry-over labialization effects will be larger than anticipatory labialization effects. Moreover, given that English /s/ is apical and therefore has higher-frequency spectral characteristics than Dutch /s/ (Quené, Orr, & Van Leeuwen, 2017), it is expected that linguistic effects will only be observed in the broadband studio recordings and not, or to a lesser extent, in the narrowband telephone recordings.

4.2 Method

4.2.1 Materials and segmentation

Materials were taken from the West Yorkshire Regional English Database (WYRED; Gold et al., 2018). This corpus contains four different speech tasks from male speakers from three different regions in Yorkshire, England. For this study, Task 2 was selected, which is a telephone conversation between a suspect (played by the participant) and an accomplice (played by a researcher in another room). Visual speech maps were used to elicit certain speech sounds. These conversations were simultaneously wiretapped from the landline telephone as well as recorded over a microphone placed in front of the participant. Participants performed Task 2 once, meaning that the within-speaker variation in this data is derived from a single 15-min telephone conversation. Since dialect was not of interest to the current study, only speakers from a single region were included, namely all 60 speakers from the Wakefield region (mean age = 21.15, *SD* = 2.85, range = 18–30).

The orthographic transcriptions that are available for each conversation were used in a forced-alignment protocol to generate segmentations at the phonemic level. To achieve the best possible

accuracy, the high-quality studio recordings were used for this. However, given the (semi-)spontaneous nature of the speech, the resulting automatic alignments were often inaccurate and needed manual correction. Target intervals were therefore estimated on four exclusion criteria and boundaries manually corrected until all speakers had at least 100 usable /s/ tokens. Tokens were excluded when they (1) were not auditorily and visually identifiable by the waveform and spectrogram as a sibilant fricative (due to reduction or elision), (2) contained interfering ambient noise or speech by the interlocutor, (3) contained laughter, or (4) contained accent imitations or other vocal imitations such as impersonations. All tokens were manually corrected and labelled on syllabic position and on whether preceding and following speech sounds were labial (consonants: /p, b, m, w/, vowels: /u, ʊ, o, ɔ, ɒ/, and (partially) rounded diphthongs: /əʊ, ɔɪ, aʊ/ were coded as labial, all other sounds were as non-labial). Diphthongs were coded as rounded irrespective of whether the rounding was immediately adjacent to the fricative (cf. temporal patterns of lip-rounding: Bell-Berti & Harris, 1982).

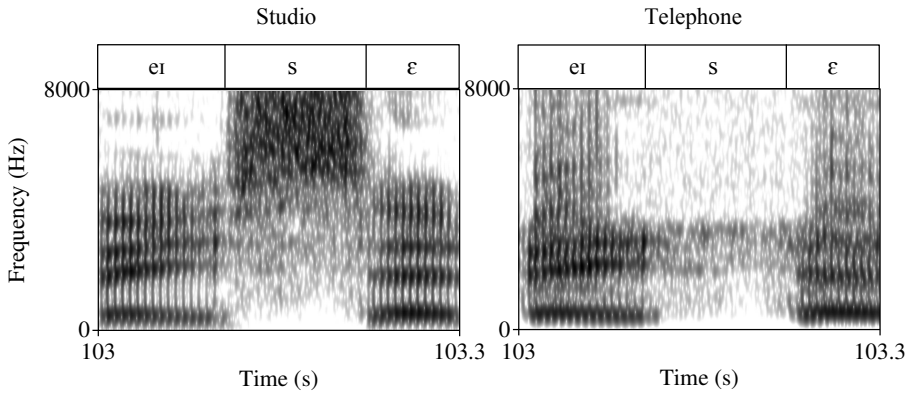

Figure 4.1: *Spectrogram for the same /s/ token in studio versus telephone channel*[6].

[6] Note that, for some tokens (such as this one), there is a very slight misalignment between the studio and telephone recording. This should only have minimal effects

For the analyses focusing on the effects of signal type and bandwidth, exactly 100 tokens per speaker ($N = 60$) were included in the analyses. For the linguistic context analysis, only speakers with at least 10 tokens per factor level were included in the analysis ($N = 55$, see Table 4.1 for the number of tokens per factor level).

Table 4.1: *Number of tokens per factor level with statistics by speaker*.

		Syllabic position		Left context		Right context	
	All	Onset	Coda	Non-labial	Labial	Non-labial	Labial
Total N	6,634	3,865	2,769	5,704	930	5,416	1,218
M	121	70	50	104	17	98	22
(*SD*)	(26)	(16)	(14)	(23)	(5)	(21)	(10)
Range	91-194	42-114	19-87	80-169	10-32	61-146	10-48

4.2.2 Acoustic analysis

Before extracting acoustic measurements for the target intervals, the simultaneously recorded telephone and studio recording for each speaker ($N = 60$) were manually aligned where needed. Given the different signal characteristics per condition, different frequency ranges were used when taking acoustic-phonetic measurements (see Table 4.2). Low frequencies up to 550 Hz were excluded to lessen the effect of ambient background

because spectral moments were measured over the middle 50% of each /s/, which is relatively stable.

noise and intruding voicing (cf. Koenig et al., 2013; Smorenburg & Heeren, 2020). For the studio condition, an upper limit of 8 kHz was chosen, because most phonetic contrasts are captured in this bandwidth in male adult speakers (although higher-frequency information plays a role in sibilants, e.g., see Monson, Lotto, & Story, 2012, the phonetic contrast between sibilants is present in the signal up to 8 kHz in male adult speakers, e.g., see Holliday, Reidy, Beckman, & Edwards, 2015). For /s/, the spectral region of interest is the one that is associated with the anterior cavity peak, found around 5 ~ 7 kHz (Koenig et al., 2013). For the studio recordings, acoustic-phonetic measurements were also taken over the 550 – 3,400 Hz range (similar to Kavanagh, 2012, see Appendix A), to see if measurements from studio and telephone recordings differed when the frequency range of measurement was equal.

Table 4.2: *Signal characteristics per channel*

	Studio	Telephone
Sampling rate [samples/s]	44,100	44,100
Frequency range [Hz]	0 – 22,050	300 – 3,400[a]
Measurement range [Hz]	550 – 8,000/3,400	550 – 3,400

[a] Telephone signal is present from 0 – 4,000 Hz, but is attenuated outside of the telephone filter of 300 – 3,400 Hz.

Four spectral moments, the spectral peak and spectral tilt were measured over the middle 50% of each /s/ token. Spectral moments capture the overall spectral shape and are often used to describe fricatives, particularly sibilants (e.g., Forrest, Weismer, Milenkovic, & Dougall, 1988; Jongman et al., 2000; Shadle & Mair, 1996). The first spectral moment (M1) is the spectral center of gravity and, in Praat (Boersma & Weenink, 2020), is computed as the mean frequency of the

spectrum in Hz. The second moment (M2) is the spectral dispersion and is computed as the variance around M1 in Hz. The third moment (L3, but M3 is also seen in the literature) is the skewness, which is a coefficient that indicates how much the spectral shape below M1 differs from that above M1, i.e., whether it leans to the left (lower frequencies) or right (higher frequencies). Lastly, the fourth moment (L4) is the kurtosis, which indicates how much the shape of the spectrum differs from a Gaussian shape, i.e., how peaked the distribution is.

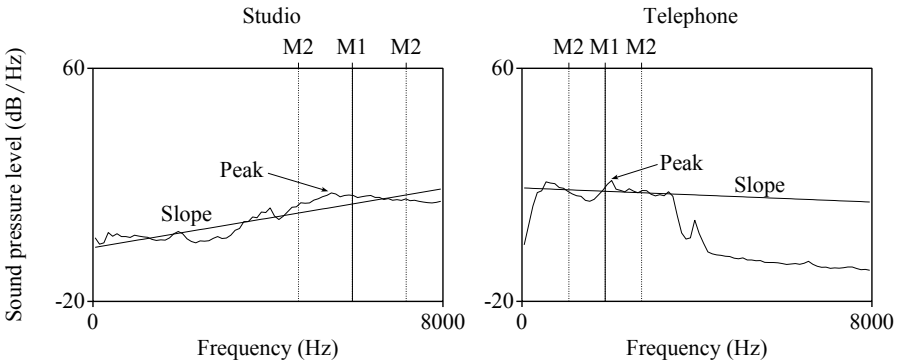

Figure 4.2: *Averaged spectra for one randomly selected speaker (WYRED speaker 041, N = 100) in the studio versus telephone channel. Measurements were taken over the 550 – 3,400 Hz range for the telephone channel and over the 550 – 8,000 Hz range for the studio channel.*

The spectral peak captures the frequency of the amplitudinal maximum in the power spectrum. For sibilant fricatives, the peak associated with the anterior cavity resonance (at 5 ~ 7 kHz: Koenig et al., 2013) falls outside of the telephone band; instead, some other amplitudinal maximum within the telephone band will be selected, which might be rather random. The spectral peak measurement should capture roughly the same type of information as M1, i.e., the size of the vocal

tract and in particular the anterior cavity, and these measurements therefore correlate highly (e.g., Ditewig et al., 2021). However, whereas the spectral peak is tied to a specific spectral event, M1 is not. M1 is highly dependent on the frequency range of measurement. L3 and L4 should also be highly affected by speech channel, as the available spectrum in the narrowband telephone filter will have a different shape than that in the broadband studio recording due to the telephone cut-offs, signal noise, and possibly the telephone codec's compression (see Figure 4.2).

Given the possible relevance of (co)articulatory information in /s/, it has been proposed that acoustic analyses of /s/ should include dynamic acoustic measurements (Koenig et al., 2013). M1 was therefore also measured dynamically, in five non-overlapping windows, each 20% of the total duration of each /s/ token. These five measurements across time were then captured in a polynomial function. Both quadratic (R^2 = 0.81, R^2 adjusted = 0.62) and cubic (R^2 = 0.92, R^2 adjusted = 0.67) functions were estimated; the cubic function was not a significantly better fit to the data than the quadratic one: $\chi^2(1) = 1.15$, $p = 0.28$. For the statistical analysis, the dynamic measures therefore consisted of two coefficients (the linear and quadratic terms). The intercept of the function was excluded because that value is conceptually the same measurement as the M1, only differing slightly in measurement window. Dynamic coefficients might be slightly more robust to speech condition because they capture the relative movement, rather than the absolute values, of M1 across the duration of /s/.

Our last measurement, the spectral tilt, refers to the overall slope of the power spectrum in the specified frequency ranges of measurement and is computed as a logarithmic regression fitted to the power spectrum using least squares. This measurement does not reflect a specific spectral event but rather a trendline of the spectrum. From Figure 4.2 it seems that the averaged spectrum in the telephone condition has a very different shape relative to the same data in the studio condition. Therefore, it is expected that, across all our data, the spectral tilt measurement is also highly affected by the telephone filter.

4.2.3 Statistical analysis

Statistical analysis was performed in R version 4.0.1. (R Core Team, 2019) and consisted of four parts. First, Pearson's correlation coefficients between acoustic-phonetic features within conditions were computed to see which features reflected the same type of information and to see which features could be combined in a follow-up speaker-discrimination test.

Second, linear mixed-effects modelling (LME) was used to firstly assess acoustic effects of the different recording types and bandwidths (Telephone 550 – 3,400 Hz versus Studio: 550 – 8,000 Hz versus Studio: 550 – 3,400 Hz) and secondly to assess the acoustic effects of Phonetic Context (NON-LABIAL, LABIAL) and Syllabic Position (ONSET, CODA) on eight acoustic-phonetic features. In the random structure of each model, a by-speaker intercept and by-speaker slopes over the fixed factors were assessed. Models were built automatically using backward stepwise elimination with BIC (Bayesian Information Criterion) estimation of random and fixed effects using function *buildmer()* from R package 'buildmer' (Voeten, 2020)[7]. The *p*-values for significance were Bonferroni-corrected for the number of acoustic measurements ($N = 8$), as several acoustic measures are extracted from the same recording and the results from these different models can therefore not be assumed to be independent.

Third, to assess speaker-specificity by recording type and bandwidth, as well as by phonetic context and syllabic position, linear discriminant analysis (LDA) was used, utilizing R package 'MASS' (Venables and Ripley, 2002). LDA is commonly used to classify a variable with multiple classes and, in speech science, is often used for

[7] Although one might expect truncated distributions for some acoustic measurements in the telephone recording, visual inspection of histograms did not show truncated distributions for any measurements. Only the spectral peak measurement showed a highly non-normal distribution, with visible peaks in the distribution at 1,500 – 2,000 Hz and 3,000 – 3,400 Hz. This indicates that, since the actual spectral peak of /s/ could not be captured due to the limited telephone band pass, other spectral peaks were found (predominantly in one of the aforementioned frequency regions).

automatic speech recognition in which speech is classified into phonetic classes (e.g., Viszlay, Juhár, & Pleva, 2012). In the current analysis, it is used to classify speakers using the acoustic-phonetic features as predictors. Speaker classifications were first computed over all data (N = 60, n = 6,000), disregarding linguistic contexts, to assess which features and combinations of features performed best at discriminating speakers and to assess the effect of the signal type and bandwidth on speaker classifications. To achieve a direct comparison, the same tokens (in each condition) were selected for the training and test data. Specifically, the first 70% of data by condition and by speaker were used as training data and the last 30% were used as test data. This way, any differences in results may be wholly attributed to signal-related effects, without potential sampling effects or other confounding variables. Before running the LDA, correlations between acoustic-phonetic features were inspected, within each of the two recording conditions. Highly-correlating predictors (r > .60) should not be entered into an LDA model together as multi-collinearity can lead to imprecise model coefficients (Klecka, 1980). The predictor set of the best-performing LDA model was used in subsequent analyses on linguistic contexts.

4.3 Results

4.3.1 Acoustic effects of the landline telephone

As expected, the M1, spectral peak, and spectral tilt measurements were highly correlated in both the studio and telephone conditions (see Table 4.3). High correlations (r > .60) were also found between M1 and L3 and between L3 and L4, although not consistently across conditions. Looking at the same measurement across conditions (see the diagonal in Table 4.3), only weak correlations were found. This suggests that the measurements in the telephone condition reflect different acoustic information than measurements in the studio condition, and also suggests large effects of the telephone filter will be found in LME modelling.

Table 4.3: *Pearson's correlations between acoustic measurements (df = 5,998) within the studio recordings (left of diagonal), within the telephone recordings (right of diagonal) and between the studio and telephone recordings (on diagonal). Significant correlations are indicated in bold.*

	Acoustic measure	Telephone M1	M2	L3	L4	$M1^{lin}$	$M1^{quad}$	Peak	Tilt
Studio	M1	**–0.44**	**–0.11**	**–0.23**	0.05	–0.03	**–0.38**	**0.78**	**0.90**
	M2	–0.05	**–0.12**	–0.07	**–0.29**	**0.13**	**–0.13**	–0.00	**–0.27**
	L3	**–0.71**	**–0.30**	**–0.08**	**0.81**	–0.02	**0.20**	**–0.23**	**–0.11**
	L4	**–0.09**	**–0.51**	**0.42**	**0.10**	–0.03	**0.08**	–0.00	**0.13**
	$M1^{lin}$	–0.02	**0.16**	–0.02	–0.05	**0.31**	0.03	–0.04	–0.06
	$M1^{quad}$	**–0.35**	**–0.23**	**0.45**	**0.14**	**0.16**	**0.13**	**–0.33**	**–0.28**
	Peak	**0.82**	–0.02	**–0.54**	–0.03	0.00	**–0.35**	**–0.26**	**0.63**
	Tilt	**0.72**	**–0.40**	**–0.15**	0.00	**–0.07**	–0.02	**0.47**	–0.01

Best-fitting LME models that assessed the effect of the telephone filter on acoustic measurements from /s/ are presented in Table 4.4. The highly-correlated measures M1, spectral peak, and spectral tilt all show large effects of the telephone filter with much lower values in the telephone than in the studio condition. According to expectations, skewness (L3) was more positive in the telephone than in the studio recording, indicating that the spectral shape is more left-leaning in the telephone condition. This makes sense, given that the telephone band has little spectral energy over 3,400 Hz. Somewhat counterintuitive, kurtosis (L4) was much higher in the telephone than the studio recording, indicating that the spectra in the telephone recording are more peaked than in the studio recording. This might be a result of the sharp cut-off of the spectrum at 3,400 Hz, resulting in a steeper peak even in the absence of the actual spectral peak (see Figure 4.2). The dynamic linear coefficient of M1 is the only measure that does not show a highly significant effect of channel, indicating that some of the dynamics of /s/

are similar across conditions. This might be related to the fact that /s/ is rather stable across time in the linear dimension. The quadratic coefficient, however, shows a large effect of condition, with a much larger dynamic movement in the studio recording. This indicates that the telephone recording does not fully capture the dynamic movement of /s/ across time.

This model was also run including a factor level for the simulation of the telephone signal, i.e., measurements taken in the studio recording using a 550 – 3,400 Hz bandwidth. Even when using this telephone-band frequency range, significant differences for all measurements (except M2, L4, and the linear coefficient of M1) were found between the studio and telephone recordings. This indicates that, although using a landline bandwidth on microphone-recorded materials makes it more similar to the landline signal, there are other differences between the conditions that are not strictly related to bandwidth.

Table 4.4: *Fixed effects in best-fitting linear mixed-effects models (N = 60, n = 6,000, default factor level = Studio: 550 – 8,000 Hz).*

	M1 [Hz]				M2 [Hz]			
Effect	*Est.*	*SE*	*t*	*p*	*Est.*	*SE*	*t*	*p*
(intercept)	5,022	74	67.8	***	1,268	12	104.6	***
Channel: Telephone	–2,943	95	30.8	***	–536	4	–135.6	***
	L3				L4			
(intercept)	0.19	0.04	4.6	***	3.85	0.99	3.9	***
Channel: Telephone	1.34	0.03	40.9	***	32.48	1.03	31.4	***

	dynamic M1^{linear} [Hz]				dynamic M1^{quadratic} [Hz]			
(intercept)	–4	26	–0.2	.8623	–739	26	–28.0	***
Channel: Telephone	81	25	3.2	.0013	545	32	17.0	***
	Peak [Hz]				Tilt [dB/decade]			
(intercept)	4,777	96	49.9	***	15.36	0.82	18.6	***
Channel: Telephone	–2,669	130	20.5	***	–11.86	1.22	9.7	***

Note. Bonferroni-corrected levels for significance: *p < 6.25e–03, **p < 1.25e–03, ***p < 1.25e–04

Regarding the random structure, best-fitting models included by-speaker intercepts for all acoustic measures. M1, spectral peak, spectral slope, and the two dynamic M1 coefficients also included by-speaker slopes over speech condition. There was a negative linear relationship between the by-speaker intercept and slope over speech condition reflecting that speakers who had higher-frequency /s/ productions showed larger acoustic effects of speech condition (see Figure 4.3), which is in line with expectations.

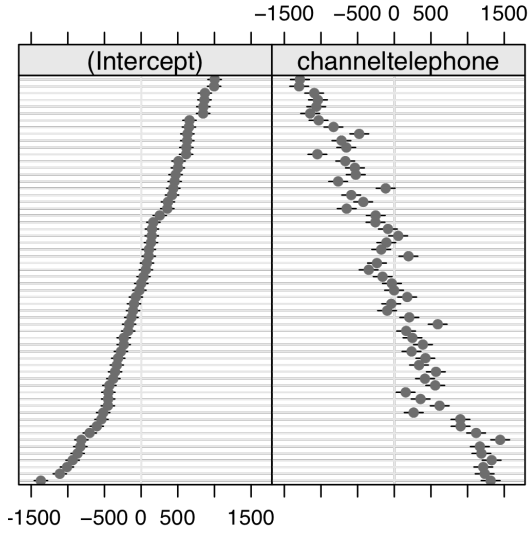

Figure 4.3: *By-speaker intercepts and slopes (N = 60) from the best-fitting LME model for M1 relative to the model intercept (5,022 Hz) and the effect of speech channel (–2,943 Hz).*

4.3.2 Acoustic effects of linguistic contexts (LME)

Starting with the measures related to the anterior resonance frequency, i.e., M1, spectral peak and tilt, these generally shows the expected effects in the studio recording. Namely, when preceding and following labial contexts or when tokens occur in coda position, the resonance frequency is lower (see Table 4.5). The effect of left context, i.e., carry-over coarticulation, is larger than that of right context, i.e., anticipatory coarticulation, which is in line with the hypothesis that English coarticulation patterns are predominantly carry-over (cf. Hoole et al., 1993). There was, however, an interaction between Right Context and Syllabic Position which showed that the effect of right labial context was larger for codas than onsets. Looking at the best-fitting models for the same speech data in the telephone recording, it can be seen that effects

are not maintained. Instead, effects in the telephone recording sometimes, but not as a rule, go in the opposite direction and generally do not resemble the patterns found in the studio recording. This indicates that detailed spectral information reflecting linguistic information is absent in the narrowband signal.

Table 4.5: *Fixed effects from linear mixed-effects models per channel*.

		Studio			Telephone		
	Effects	*Est.*	*SE*	*t*	*Est.*	*SE*	*t*
M1	(intercept)	5,190	77	67.3	2,075	32	64.2
[Hz]	Left context = LABIAL	–365	20	–18.7	112	10	10.6
	Right context = LABIAL	–94	22	–4.3	–31	12	–2.6
	Syll. Position = CODA	–200	15	–13.2	–1	8	–0.1
	Right x Syll. Position	–118	37	–3.2	68	20	3.4
M2	(intercept)	1,249	26	48.7	728	6	123.5
[Hz]	Left context = LABIAL				21	4	4.8
	Right context = LABIAL				25	4	6.2
	Syll. Position = CODA	35	6	6.0			
L3	(intercept)	0.04	0.09	0.5	1.58	0.07	21.9
	Left context = LABIAL	0.48	0.03	17.3	–0.33	0.08	–4.2
	Right context = LABIAL	0.19	0.03	7.4			
	Syll. Position = CODA	0.13	0.02	6.5	–0.19	0.06	–3.4
L4	(intercept)	3.75	0.37	10.2	34.63	1.50	23.2
	Left context = LABIAL	0.75	0.22	3.4			
Tilt	(intercept)	17.0	0.8	21.2	3.9	0.9	4.5
[dB/	Left context = LABIAL	–2.3	0.2	–10.7	2.0	0.3	6.6
decade]	Right context = LABIAL	–0.5			–1.9	0.3	–5.6
	Syll. Position = CODA	–2.2	0.2	–13.3	–0.6	0.2	–2.4
	Right x Syll. Position	–1.5	0.4	–3.6	2.5	0.6	4.3
$M1^{lin}$	(intercept)	–173	29	–5.9	17	17	1.0
[Hz]	Left context = LABIAL	144	33	4.4			
	Right context = LABIAL	–259	37	–7.0	–22	15	–1.5
	Syll. Position = CODA	492	25	19.4	136	10	13.3
	Right x Syll. Position	–276	62	–4.4	119	25	4.8
$M1^{quadr}$	(intercept)	–761	29	–26.1	–194	12	–16.7
[Hz]	Right context = LABIAL	–70	22	–3.2	–31	8	–3.8
	Syll. Position = CODA	116	15	7.7	26	7	3.8
	Left x Syll. Position				–105	19	–5.7
	Right x Syll. Position	185	37	–5.0			

4.3.3 Telephone effects on speaker discrimination (LDA)

In Table 4.6, the speaker-classification accuracies for the different LDA models are presented. With 60 speakers, chance level for classification accuracy was 1/60 (= 1.7%). Amongst the individual measures, the ones associated with the size of the anterior cavity, i.e., M1, spectral peak and tilt, performed best at discriminating speakers. M1 reached the highest accuracy, but spectral tilt was more robust to the telephone filter, possibly because spectral tilt – as a trend line fitted to the spectrum – is less tied to specific spectral events than M1. It seems that M1, spectral peak and tilt contain the most idiosyncratic information. Regarding the effect of condition, as expected, acoustic measures taken in the studio recording generally have more discriminatory power than acoustic measures taken in the telephone recording. The spectral tilt and the linear and quadratic terms of M1 showed only minor differences between speech channels and thus seem to be the most robust to bandwidth effects. Following the acoustic results, this was expected for the linear term; smaller acoustic effects should correspond to smaller effects of condition on the speaker classification. Despite the fact that spectral tilt and the quadratic term showed significant bandwidth effects on the acoustics, they seem relatively robust to these effects on the speaker classification, as we initially predicted based on the nature of these measurements.

Table 4.6: *LDA speaker-classification accuracies (in %) for independent features and combined features across recording type and bandwidth.*

Predictor (set)	Studio (550 - 8,000 Hz)	Telephone (550 - 3,400 Hz)	Studio (550 - 3,400 Hz)
M1	5.7	3.9	5.2
M2	4.1	2.9	3.3
L3	4.4	2.1	2.0
L4	3.1	2.0	2.4
$M1^{linear}$	2.9	3.0	1.9
$M1^{quadratic}$	3.1	2.9	2.3
peak	4.7	3.2	2.9
tilt	4.8	4.4	5.1
M1 + M2 + L4	9.7	5.6	6.2
M2 + L3	7.8	2.8	3.9
$M1^{linear}$ + $M1^{quadratic}$	4.2	2.9	2.5
M1 + M2 + L4 + $M1^{linear}$ + $M1^{quadratic}$	11.4	6.7	6.8
M2 + L3 + $M1^{linear}$ + $M1^{quadratic}$	9.4	4.2	4.0
M2 + L3 + tilt	12.9	6.5	6.8
M2 + L3 + $M1^{linear}$ + $M1^{quadratic}$ + tilt	14.5	7.5	7.6

The best-performing LDA model, with M2, L3, the linear and quadratic M1 coefficients and the spectral tilt as predictors had a 14.5% accuracy (95% CI [12.5, 15.8]) for the studio data and a 7.5% accuracy (95% CI [6.5, 8.8]) for the telephone data. For both the studio and telephone data, two linear discriminant (LD) functions were needed to account for ~80% of the between-speaker variance (studio: LD1 = 48%, LD2 = 31%; telephone: LD1 = 66%, LD2 = 15%). Mirroring the results on the models with individual predictors, the scaling coefficients from this combined model further indicate that LD1 in both the studio and

telephone data was largely explained by the spectral tilt, i.e., this was the most-contributing predictor. For LD2, L3 was the most-contributing predictor. The scaling coefficients also indicated that the dynamic linear and quadratic terms of M1 had the least discriminatory power, which is in accordance with previous research on /s/ in Dutch spontaneous speech (Smorenburg & Heeren, 2020). In spontaneous speech, /s/ dynamics are probably mainly determined by contextual effects, leaving little room for idiosyncrasies in articulatory dynamics (cf. Heeren, 2020b).

In the acoustic analysis, strong positive correlations were found between by-speaker intercepts and slopes over speech channel for M1, peak, tilt, and the dynamic M1 coefficients. This indicates that speakers with high average values for measurements also showed larger effects of speech channel. Despite this, there does not seem to be a strong relationship between the size of speech channel effects on the acoustics and on the speaker-classification accuracy on the speaker level. Pearson's correlations between speakers' channel effects on the acoustics and speaker-classification accuracy from the best-performing LDA model were not significant for any of the measures except for a weak correlation for spectral slope ($r = -.26$, $p < .05$).

As can be seen in Table 4.7, the speaker-classification accuracy is much better in the studio than in the telephone recording across all conditions. Linguistic context does not affect the speaker-classification accuracy save one exception: when the right phonetic context is labial, there is better performance than when the right phonetic context is non-labial. However, in the telephone recording, this difference is neutralized. All other differences between contexts are considered negligible because they are smaller than chance level accuracy (100% / 55 speakers = 1.82%).

Table 4.7: *LDA speaker-classification accuracies (in %) per factor level and recording type using the predictors from the best-performing model from Table 4.6.*

	Context	Studio (550 - 8,000 Hz)	Telephone (550 – 3,400 Hz)
All		14.5	7.5
Syllabic position	ONSET	14.4	7.6
	CODA	14.7	8.2
Left context	LABIAL	14.9	7.1
	NON-LABIAL	14.2	6.7
Right context	LABIAL	20.2	6.2
	NON-LABIAL	14.5	6.7

4.4 Discussion

Previous research has found that vowels' F1 measurements in telephone signals may shift upward by an average of 15% in landline signals and by an average of 29% (with up to 60% rises in F1 values) in mobile signals relative to studio recordings (Künzel, 2001; Byrne & Foulkes, 2004). As expected, because the sibilant fricative's spectral peak falls far outside of the upper limit of the narrowband telephone filter, the effect of landline filters on sibilant fricative /s/ acoustics is much larger than that of F1 for several vowels. This is, of course, mostly a reflection of the decrease in bandwidth in the telephone channel relative to the studio channel. Whereas the telephone filter only shaves off some of the spectral energy for F1, the average spectral peak for /s/ was 4,777 Hz in the studio recording (i.e., 1,377 Hz above the telephone band's upper limit), making it impossible to measure in the telephone signal.

These large effects of speech condition on the acoustics were reflected in the idiosyncratic information in /s/ as shown by the speaker

classification results; in the best-performing model, classification accuracy decreased by about half as a function of bandwidth. From these results we can conclude that the signal from 550 – 3,400 Hz does not capture much of the between-speaker variation that is present in the 550 – 8,000 Hz signal. This is in accordance with the observation that fricative discrimination in broadband signals is better than in narrowband signals (Bessette et al., 2002). However, the current results also show that some idiosyncratic information remains in /s/ from landline telephone speech, as speaker-classification accuracies on telephone speech are (at least slightly) above chance level in the LDA results (see also Smorenburg & Heeren, 2020).

Furthermore, for some acoustic measures, findings from the correlations, the acoustic analysis, and the speaker-classification analysis show interesting patterns. Spectral tilt, e.g., shows no correlation across channels ($r = -.01$, see Table 4.3), which is congruent with the acoustic analysis that shows large effects of speech condition on this measurement. In speaker classification, however, spectral tilt performs relatively well, with only a minor difference (0.4%) in classification accuracy between bandwidths. This implies that, while the measured tilt is significantly altered (lower in the telephone channel), the same amount of speaker information is available in the measurement. Another interesting observation is that the M1 and spectral peak measurements are highly correlated within recording types, even though the spectral peak that is usually targeted (often the spectral maximum around 5 ~ 7 kHz, Koenig et al., 2013) cannot be measured accurately in the telephone band. We expected that the peak measurement in the telephone recordings would therefore be rather random. However, its high correlation with the M1 measurement and the above-chance speaker-classification accuracy in the telephone recording seem to indicate that spectral peak measurements in telephone recordings still systematically capture some information about resonance properties in /s/. This is further corroborated by the distribution of spectral peak measurements, which shows a bimodal distribution (see footnote 2). This indicates that, when the actual spectral peak cannot be measured in the telephone band, another peak is found, not randomly, but predominantly in one of two specific frequency regions.

As for linguistic sampling context, British English /s/ acoustics show effects of contextual labialization and coda reduction in the studio recording, including an interaction between anticipatory labialization and syllabic context which showed more anticipatory labialization for codas. Interestingly, we find evidence for the hypothesis that English has predominantly carry-over coarticulation (Hoole et al., 1993), as effects of left context are larger than effects of right context, i.e., anticipatory coarticulation. In earlier work on Dutch, /x/ showed somewhat larger anticipatory coarticulation, also particularly in coda position (Smorenburg & Heeren, 2020). This might be indicative of other labial coarticulation patterns in British English versus Dutch, with the former being more carry-over and the latter more anticipatory in nature. Comparing linguistic effects on the acoustics across recording type, we see that acoustic effects are not maintained in the telephone recording. In fact, effects in the landline telephone recording are not similar to effects observed in the studio recording and are also not congruent with previous findings on linguistic effects on fricative acoustics; they do not seem to follow any discernable pattern relative to effects in the same speech data recorded over the studio recording. Remember that these results were obtained using landline telephone signals. Although landlines are still in use and therefore relevant, mobile signals are common in daily communications. Mobile signals differ from landline signals in that they can have varying bit rates and therefore varying bandwidths. Future work should consider also examining the effects of mobile signals on different speech sounds across linguistic contexts.

The linguistic effects generally seemed to have no effect on the amount of speaker information in /s/, with the exception of one phonetic context: /s/ tokens perform better when followed by labial segments, which is a context with increased between-speaker variation assumedly due to speaker-specific patterns in anticipatory labialization. This effect was only observable in the studio recording and seemed to be neutralized in the telephone recording.

Previous research has shown that listeners are generally less able to identify speakers over the telephone than over studio-recorded speech (Reynolds, 1995). Knowing that /s/ is a relatively speaker-specific consonant, the results of the current study may contribute to explaining

why speaker identification is lower in telephone speech. When looking at the segmental level, it has also been shown that certain acoustic-phonetic features may be associated with certain social factors. For /s/, some of its acoustic-phonetic features have been linked to social factors regarding gender and sexual orientation (Munson et al., 2006; Tracy et al., 2015). Although other acoustic-phonetic features (from vowels) also encode this type of information, it would be interesting to see whether the telephone effect on individual acoustic-phonetic features affects the perception of social information. For example, can listeners perceive a speaker's sexual orientation equally well from narrowband telephone signals as from broadband signals? Previous research has identified spectral skewness (L3) as an important feature in the perception of sexual orientation of male speakers (Munson et al., 2006). The current results show large effects of speech channel on both the acoustics and speaker classification of L3, which might mean that the perception of sexual orientation from /s/ is more difficult in telephone speech.

4.5 Conclusion

To conclude, for forensic speech science, it seems clear that the idiosyncratic information contained in telephone speech is severely compromised compared to studio-recorded speech. The current analysis on /s/ represents an extreme case of the telephone effect due to the high-frequency spectral characteristics of /s/; the telephone filter causes large changes in acoustic-phonetic measurement values and in speaker-classification accuracies. Despite large acoustic effects of speech channel, some measurements, in particular spectral tilt, showed relatively small effects of speech channel on speaker classification and can therefore still be useful for speaker discrimination in telephone speech. As for linguistic sampling context, although the landline telephone filter greatly affects the presence of expected linguistic effects on the acoustics, these are generally not, or only slightly, reflected in LDA speaker-classification accuracy.